ED 220 474 TM 820 369

AUTHOR Kolen, Michael J.; Hoover, H. D.
TITLE The Reliability of Selected Item Bias Procedures.
PUB DATE Mar 82
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Black Students; *Elementary School Students; Females; *Grade 5; Latent Trait Theory; Males; Racial Differences; *Reliability; Sex Differences; *Test Bias; *Test Items; White Students
IDENTIFIERS *Iowa Tests of Basic Skills

ABSTRACT

The reliabilities of 6 item bias indices and correlations among them were investigated for each of the 11 tests of the Iowa Tests of Basic Skills (ITBS). The difficulty and delta indices detected group differences in relative item difficulty. Biserial and point biserial indices detected group differences in item discrimination. The Scheuneman and three-parameter indices detected differences in relative item difficulty by score level and latent ability level, respectively. The sample group consisted of 800 fifth grade students evenly divided between black and white, male and female students, thus examining racial and sexual bias. The results indicated the item bias indices investigated were unreliable when based on sample sizes of 200 minority and 200 majority examinees. The instability of the indices may have resulted from the fact that very few, if any, biased items are included in the ITBS. The study suggested that the use of item bias indices to screen achievement test items could not be expected to lead to consistent decisions about which items are biased with small sample sizes. (DWH)

# The Reliability of Selected Item Bias Procedures

Michael J. Kolen

American College Testing Program

H. D. Hoover

The University of Iowa

Key Words: Item Bias
Latent Trait
Reliability

Paper presented at the Annual Meeting of the American Educational Research Association in New York City, March, 1982

# The Reliability of Selected Item Bias Procedures

Michael J. Kolen

American College Testing Program

H. D. Hoover

The University of Iowa

## Abstract

The reliabilities of six item bias indices are investigated for each of the eleven tests of the Iowa Tests of Basic Skills, using random samples of fifth grade students. Both racial and sexual bias are considered. The reliability of an index is defined here as its stability from one randomly equivalent group to another. The results indicate that the item bias indices investigated are fairly unreliable when based on sample sizes of 200 minority and 200 majority examinees. Consequently, this study suggests that the use of item bias indices to screen achievement test items cannot be expected to lead to consistent decisions about which items are biased with sample sizes of about 200. Additionally, correlations among bias indices are investigated.

# The Reliability of Selected Item Bias Procedures

The elimination of biased (cultural, racial, sexual, etc.) items from achievement tests is often conceptualized to be a two-stage process. First, "experts" judge the fairness of the presentation format and content of the items for a variety of groups. Those items which are judged to be unfair, or biased, are excluded from the test. Second, many researchers, including Scheuneman (1979), have advocated the use of item bias statistics to screen test items prior to the construction of final test forms. Ideally, bias indices would be calculated from item tryout data. Based on these indices, biased items would be excluded from the test in much the same way that test items with low item discriminations are excluded in the item tryout stages of test development.

Item bias indices should produce stable results if they are to be used beneficially for screening purposes. However, certain studies suggest that item bias statistics may be fairly unstable. Studies by Plake (1980) and Qualls and Hoover (1981) suggested that the statistical bias indices are only minimally related to "experts'" judgments of item bias. Scheuneman (1980) and Linn, Levine, Hastings and Wardrop (1981) found only modest agreement among item bias statistics across independent samples. Linn et al. (1981) concluded that ". . . it may be difficult to identify biased items because of the unreliability of the indices used" (p. 170).

None of the previously completed studies directly addressed the issue of the reliability (stability from one randomly equivalent group to another) of item bias indices. For this reason, the reliabilities of each of six internal criterion item bias indices were investigated in the present study. Indices were calculated for both race and sex categorizations for each of the eleven tests of the Iowa Tests of Basic Skills administered to fifth grade students. Only unsigned versions of the indices (Ironson & Subkoviak, 1979) were investigated since item screening, as usually conceived, involves eliminating items biased against any group. The indices were based on samples of 200 examinees from each race or sex categorization. These sample sizes were viewed as being the largest which typically would be available for minority students in most item tryout situations. Additionally, the relationships among item bias indices were examined.

No discussion of the differences among definitions of item bias or among item bias statistics will be presented here. These issues are discussed in a variety of sources including Hunter (1975), Ironson and Subkoviak (1979), Lord (1980), Marascuilo and Slaughter (1981), Rudner, Getson, and Knight (1980a,b); and Shepard, Camilli, and Averill (1981).

## Item Bias Indices

Six different item bias indices were evaluated in this study. The difficulty and delta indices to be discussed were designed to detect group differences (e.g., between blacks and whites) in relative item difficulty. The biserial and point biserial indices were designed to detect group differences in item discrimination. The Scheuneman and 3-parameter indices were designed to detect differences in relative item difficulty by score level and latent ability level, respectively.

## Difficulty and Delta Indices

The *difficulty* index was referred to as the transformed item difficulties--45° line method by Rudner et al. (1980a), except that the absolute value of the Rudner et al. (1980a) index was used in the present study. For this index, item difficulties (p-values) are calculated and standardized (mean of zero; standard deviation of one) within each group. The *difficulty* index for an item is calculated as the absolute value of the difference between standardized item difficulty for the two race or sex groups.

The *delta* index was referred to as the transformed item difficulties--major axis index by Rudner et al. (1980a) with one substantive modification--the *delta* index is the absolute value of the Rudner et al. (1980a) index. For this index, the within group item difficulties are transformed using the inverse normal transformation. These transformed difficulties are then standardized (mean of zero; standard deviation of one) within groups. The *delta* index for an item is the absolute difference between the standardized transformed difficulties for the two groups. A similar approach was used by Angoff and Ford (1973).

## Biserial and Point Biserial Indices

The *biserial* index for an item is the absolute difference between the within group biserial correlations of the item with total score. The *point biserial* index for an item is the absolute difference between the within group point biserial correlations of item with total score.

## Scheuneman Index

The *Scheuneman* index (Scheuneman, 1979) was calculated for each item using five score levels. The score levels were defined such that approximately equal numbers of examinees were in each level. According to Scheuneman (1979), the index could be expected to be distributed approximately chi-square with four degrees of freedom.

3-Parameter Index

The 3-parameter index is a modification of the index proposed by Linn and Harnisch (1981). This index was chosen because it can be used with smaller sample sizes than the more widely recommended index suggested by Lord (1980). For this index, first the item and ability parameters of the three-parameter logistic item response theory model are estimated for the combined group of examinees. For example, item responses for black and white students are pooled in order to estimate the model parameters. The two groups of examinees are then separated. For each examinee, the difference between the examinee's estimated probability (p) of correctly answering the item and the examinee's actual response to the item (1=correct; 0=incorrect) is found. This quantity is then divided by a standard error--p(1-p)--and averaged over examinees within each group. The mean for each group is then squared and the two squared means summed to arrive at the 3-parameter index.

## Method

The data consisted of item responses by 800 fifth grade students who participated in the 1977 national standardization of the Iowa Tests of Basic Skills (ITBS). The sample included 200 black males, 200 black females, 200 white males, and 200 white females with equal numbers of each of these groups randomly selected from individual schools in the standardization sample. Thus, the sample contained equal numbers of black and white pupils and was balanced by sex. In addition, the confounding of curriculum differences and ethnic group membership, common to many item bias studies, was partially controlled. All eleven tests from the ITBS were analyzed.

The black students were randomly divided, stratified by sex, into two samples of 200 students each. The same procedure was followed for white students. Item bias statistics were calculated for the first sample of black vs. the first sample of white students as well as for the second sample of black vs. the second sample of white students. The item bias indices were calculated separately for each of the eleven ITBS tests. Identical procedures were followed for the female vs. male comparisons except that the stratification in the random sampling was by race.

The reliability of each item bias index was investigated by test for the race categorization as well as for the sex categorization. The correlation between the values of an item bias index across random samples was used as a measure of the reliability of the index. Additionally, items were classified as either biased or unbiased using the _difficulty_, _delta_, and _Scheuneman_ indices. Items with _difficulty_ or _delta_ indices above 0.75 were classified as biased by that index on the suggestion of Rudner et al. (1980b). Items with _Scheuneman_ index values which surpassed the 0.05 critical value of a chi-square distribution with four degrees of freedom were classified as biased on the recommendation of Scheuneman (1979). The agreement in classification of items across random samples by a given index was used as another method to investigate the reliabilities of each of these three item bias indices.

The values of each item bias index were pooled over all of the items in the test battery and the reliability of each index and the intercorrelations among indices--across randomly equivalent samples--were estimated. Additionally, disattenuated intercorrelations were estimated in order to investigate the relationships among item bias statistics in the presence of no estimation error.

## Results

An attempt was made to estimate the three-parameter logistic item response model parameters using separate LOGIST (Wood, Wingersky, and Lord, 1978) runs for each randomly equivalent sample of 400 examinees. However, LOGIST failed to converge. Because of these convergence problems, the parameter estimation was completed using all 800 examinees. The 3-parameter indices were calculated using these parameter estimates following the same general procedures as were followed for the other indices. The use of parameter estimates from the combined sample results in a dependency between indices across randomly equivalent samples. Therefore, the reported reliabilities for the 3-parameter index are probably overestimates of the actual values of the index. For this reason, the index was calculated only for the vocabulary and language usage tests of the ITBS.

The means and standard deviations of raw scores on each test are presented in Table 1. The means and standard deviations were generally larger for whites than for blacks. There also appeared to be a tendency for the females in this sample to earn slightly higher scores than the males.

--------------------------------

Insert Tables 1 and 2 about here

--------------------------------

The reliabilities of item bias indices for the race comparison are presented in Table 2. Very few of the reliabilities surpassed the .05 critical value. The reliabilities were generally in the very low to, at best, moderate range. The reliabilities for the language usage test were the only ones which were consistently moderate across indices. Overall, the difficulty and delta indices tended to produce more reliable results than any of the other indices for the race comparison. However, Hunter (1975) illustrates how mean differences between groups can lead to large values of these bias statistics, even when the item is not biased. Thus, the reliability of these indices may have been more of an artifact of the substantial mean differences between blacks and whites than reliability for detecting item bias, per se. Additionally, the Scheuneman index tended to produce the least reliable results for the race comparison. Also, note that for the vocabulary and language usage tests, the 3-parameter index tended to have a lower reliability than the other indices.

The reliabilities of the item bias indices for the sex comparison are presented in Table 3. The reliabilities were generally very low. In fact, there is little evidence to suggest that the reliabilities for any index, except possibly the Scheuneman index, were above zero.

Note that reliabilities of signed indices are included in the Appendix for the sake of completeness. Tables corresponding to Tables 2 and 3 are provided.

---

Insert Tables 3 and 4 about here

---

The intercorrelations among item bias indices across all tests for the race comparison are shown in Table 4. The diagonal entries represent the indices' reliabilities across tests. These reliabilities were fairly low.

The values above the diagonal represent the average intercorrelations among indices across samples. For example, the 0.29 value in the table represents the average of two correlations. The first was the correlation between the difficulty index for the first random sample and the delta index for the second random sample. The second correlation included in the average was between the difficulty index for the second random sample and the delta index for the first random sample. The values above the diagonal were used in combination with the reliabilities to arrive at the disattenuated correlations presented below the diagonal in Table 4.

The disattenuated correlations strongly suggest that the difficulty and delta indices both reflect the same item bias property and that the biserial and point biserial indices both reflect the same item bias property. The disattenuated correlations also strongly suggest that the difficulty and delta indices reflect a very different item bias property than that reflected by the biserial and point biserial indices. Additionally, the disattenuated correlations suggest that the Scheuneman index reflects properties reflected by both the difficulty/delta indices and biserial/point biserial indices of item bias.

Table 5 presents the intercorrelations among bias indices for the sex comparison. The reliabilities as well as the intercorrelations among indices were negligible. Disattenuated correlations are not presented as all of the reliabilities in the table failed to surpass the .05 critical value. Overall, the results suggested little or no consistency for the sex comparison across random samples, for any index.

---

Insert Table 5 about here

---

The numbers of items classified as biased by the difficulty, delta, and Scheuneman indices are presented in Table 6 for the race comparison and in Table 7 for the sex comparison. The results presented suggest that there was minimal agreement across randomly equivalent samples, at best.

--------------------------------

Insert Tables 6 and 7 about here

--------------------------------

## Discussion

The results suggested that the item bias indices investigated are fairly unreliable when based on sample sizes of 200 minority and 200 majority examinees. The use of item bias indices to screen achievement test items for bias could not be expected to lead to consistent decisions about which items are biased with these sample sizes.

One potential explanation of the instability of the indices is that few, if any, biased items are included on the ITBS. In the ITBS test construction procedures, the content and presentation format of the test items are evaluated for bias using "experts'" judgments. Perhaps, the use of the judgments of "experts" is sufficient to detect biased items in achievement tests and the item bias statistics provide little additional information. If so, then it would be more beneficial for test constructors to use available resources to hire "experts" to screen items rather than to compute item bias indices.

If the screening of items for bias using item bias indices is to produce beneficial results, then research is needed to ascertain the sample sizes necessary to produce sufficiently stable results. The present study clearly showed that sample sizes of 200 minority and 200 majority examinees are too small to allow for reliable decisions of bias based on the bias indices that were investigated.

## References

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 1973, *10*, 95-105.

Hunter, J. E. *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items.* Paper presented at the National Institute of Education Invitational Conference on test bias, Annapolis, 1975.

Ironson, G. H., & Subkoviak, M. J. A comparison of several methods of assessing item bias. *Journal of Educational Measurement*, 1979, *16*, 209-225.

Linn, R. L., & Harnisch, D. L. Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, 1981, *18*, 109-118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 1981, *5*, 159-173.

Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale, N.J.: Erlbaum, 1980.

Marascuilo, L. A., & Slaughter, R. E. Statistical procedures for identifying possible sources of item bias based on $\chi^2$ statistics. *Journal of Educational Measurement*, 1981, *18*, 229-248.

Plake, B. S. A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement*, 1980, *40*, 397-404.

Qualls, A., & Hoover, H. D. *Black and white teacher ratings of elementary achievement test items for potential race favoritism.* Paper presented at the Annual Convention of the American Educational Research Association, Los Angeles, April 1981.

Rudner, L. M., Getson, P. R., & Knight, D. L. A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, 1980a, *17*, 1-10.

Rudner, L. M., Getson, P. R., & Knight D. L. Biased item detection techniques. *Journal of Educational Statistics*, 1980b, 213-233.

Scheuneman, J. A new method for assessing bias in test items. *Journal of Educational Measurement*, 1979, *16*, 143-152.

Scheuneman, J. *Consistency across administrations of certain indices of bias in test items.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, 1980.

Shepard, L., Camilli, G., & Averill, M. Comparison of six procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 1981, *6*, 317-375.

Table 1

Mean and Standard Deviation[a] of Raw Scores by Race and Sex

| | Number of items | Race | | Sex | | Overall |
|---|---|---|---|---|---|---|
| | | Blacks | Whites | Females | Males | |
| Vocabulary | 39 | 13.61<br>( 6.93) | 21.66<br>( 9.22) | 17.87<br>( 8.65) | 17.40<br>( 9.51) | 17.63<br>( 9.09) |
| Reading | 54 | 17.46<br>( 7.13) | 26.30<br>(11.02) | 22.23<br>( 9.81) | 21.53<br>(10.72) | 21.88<br>(10.27) |
| Spelling | 40 | 17.46<br>( 8.86) | 22.16<br>( 9.28) | 22.02<br>( 9.16) | 17.90<br>( 9.05) | 19.96<br>( 9.33) |
| Capitalization | 30 | 12.31<br>( 4.69) | 15.83<br>( 5.72) | 15.00<br>( 5.29) | 13.13<br>( 5.58) | 14.07<br>( 5.51) |
| Punctuation | 30 | 10.60<br>( 4.60) | 14.70<br>( 6.16) | 13.52<br>( 5.85) | 11.78<br>( 5.63) | 12.65<br>( 5.80) |
| Language Usage | 30 | 9.60<br>( 4.74) | 15.50<br>( 6.75) | 13.15<br>( 6.44) | 11.95<br>( 6.58) | 12.55<br>( 6.53) |
| Visual Materials | 46 | 15.59<br>( 5.29) | 21.84<br>( 7.51) | 18.65<br>( 6.64) | 18.77<br>( 7.74) | 18.71<br>( 7.21) |
| Reference Materials | 45 | 17.57<br>( 7.21) | 23.82<br>( 9.71) | 21.90<br>( 8.91) | 19.49<br>( 9.15) | 20.69<br>( 9.10) |
| Math Concepts | 37 | 12.97<br>( 5.30) | 17.71<br>( 6.64) | 15.65<br>( 6.19) | 15.03<br>( 6.70) | 15.34<br>( 6.46) |
| Math Problem Solving | 27 | 9.54<br>( 4.15) | 13.10<br>( 5.41) | 11.27<br>( 4.80) | 11.36<br>( 5.46) | 11.32<br>( 5.14) |
| Math Computation | 45 | 19.85<br>( 7.43) | 22.32<br>( 8.18) | 22.24<br>( 7.64) | 19.92<br>( 8.01) | 21.08<br>( 7.91) |

[a] Numbers in parentheses represent standard deviations.

Table 2

Reliability of Item Bias Indices for Race

| Test | Number of Items | Bias Index | | | | | |
|---|---|---|---|---|---|---|---|
| | | Difficulty | Delta | Biserial | Point Biserial | Scheuneman | 3-Parameter[a] |
| Vocabulary | 39 | .38* | .32* | .22 | .43* | .06 | .25 |
| Reading | 54 | .25 | .19 | .04 | .18 | -.16 | |
| Spelling | 40 | .24 | .21 | .04 | .08 | .24 | |
| Capitalization | 30 | -.09 | -.07 | .44* | .47* | .31 | |
| Punctuation | 30 | .45* | .35* | .17 | .24 | .26 | |
| Language Usage | 30 | .48* | .55* | .49* | .64* | .55* | .36* |
| Visual Materials | 46 | .41* | .24 | .07 | .18 | .04 | |
| Reference Materials | 45 | .01 | -.07 | .03 | .07 | .06 | |
| Math Concepts | 37 | .19 | .14 | .21 | .14 | -.30 | |
| Math Problem Solving | 27 | .13 | .08 | .29 | .37* | .04 | |
| Math Computation | 45 | -.06 | -.01 | .04 | .09 | .30 | |
| Median | k | .24 | .19 | .17 | .18 | .06 | -- |

* p < .05

[a] Index was computed only for those tests with values in this column.

Table 3

Reliability of Item Bias Indices for Sex

| Test | Number of Items | Bias Index | | | | | |
|---|---|---|---|---|---|---|---|
| | | Difficulty | Delta | Biserial | Point Biserial | Scheuneman | 3-Parameter[a] |
| Vocabulary | 39 | .22 | .22 | .14 | .11 | .01 | .09 |
| Reading | 54 | .19 | .15 | .08 | .12 | .34* | |
| Spelling | 40 | -.19 | -.15 | -.23 | -.23 | .00 | |
| Capitalization | 30 | -.21 | -.13 | -.14 | -.19 | .18 | |
| Punctuation | 30 | .23 | .19 | -.15 | -.11 | .11 | |
| Language Usage | 30 | -.11 | -.14 | -.05 | -.03 | .31 | -.16 |
| Visual Materials | 46 | .14 | .10 | .21 | .18 | .03 | |
| Reference Materials | 45 | -.15 | -.16 | -.09 | -.09 | .08 | |
| Math Concepts | 37 | -.04 | -.10 | .22 | .22 | .37* | |
| Math Problem Solving | 27 | -.17 | -.12 | .05 | .07 | -.02 | |
| Math Computation | 45 | .10 | .04 | .00 | -.11 | .38* | |
| Median | | -.04 | -.10 | .00 | -.03 | .11 | |

* p < .05.

[a] Index was computed only for those tests with values in this column.

Table 4

Correlations Between Item Bias Indices Across All

Tests for Race

| Index | Difficulty | Delta | Biserial | Point Biserial | Scheuneman |
|---|---|---|---|---|---|
| Difficulty | .34* | .29* | .00 | .01 | .06 |
| Delta | .99+ | .27* | .02 | .03 | .07 |
| Biserial | .01 | .08 | .22* | .26* | .11* |
| Point Biserial | .01 | .11 | .99 | .32* | .11* |
| Scheuneman | .45 | .36 | .59 | .51 | .15* |

* p < .05

Note: Diagonal values are reliabilities across all tests. Values above the diagonal are average correlations between indices across all tests. Values below the diagonal are disattenuated correlations between indices across all tests. Correlations were based on 423 items.

Table 5

Correlations Between Item Bias Indices Across All Tests for Sex

| Index | Difficulty | Delta | Biserial | Point Biserial | Scheuneman |
|---|---|---|---|---|---|
| Difficulty | 0.08 | 0.07 | 0.00 | 0.01 | 0.01 |
| Delta | | 0.07 | 0.01 | 0.03 | 0.03 |
| Biserial | | | 0.03 | 0.02 | 0.08 |
| Point Biserial | | | | 0.02 | 0.07 |
| Scheuneman | | | | | 0.07 |

Note: None of the correlations surpassed the .05 critical value. Diagonal values are reliabilities across tests. Values above the diagonal are average correlations between indices across all tests. Correlations were based on 423 items.

Table 6

Number of Biased Items for Race

| Test | Number of Items | Index | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Difficulty | | | Delta | | | Scheuneman | | |
| | | Number of Biased Items | | | | | | | | |
| | | Sample One | Sample Two | Both Samples | Sample One | Sample Two | Both Samples | Sample One | Sample Two | Both Samples |
| Vocabulary | 39 | 3 | 6 | 0 | 3 | 5 | 0 | 1 | 1 | 0 |
| Reading | 54 | 2 | 5 | 0 | 2 | 3 | 0 | 0 | 1 | 0 |
| Spelling | 40 | 1 | 5 | 0 | 1 | 4 | 0 | 0 | 0 | 0 |
| Capitalization | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 |
| Punctuation | 30 | 4 | 6 | 1 | 5 | 8 | 2 | 1 | 0 | 0 |
| Language Usage | 30 | 6 | 4 | 2 | 6 | 4 | 2 | 1 | 1 | 1 |
| Visual Materials | 46 | 8 | 7 | 3 | 5 | 4 | 1 | 0 | 1 | 0 |
| Reference Materials | 45 | 5 | 2 | 1 | 6 | 3 | 1 | 0 | 0 | 0 |
| Math Concepts | 37 | 2 | 1 | 0 | 3 | 1 | 0 | 1 | 0 | 0 |
| Math Problem Solving | 27 | [illegible] | 3 | 0 | 2 | 3 | 0 | 1 | 2 | 0 |
| Math Computation | 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| Overall | 423 | 33<br>( 7.8%) | 39<br>( 9.2%) | 7<br>(1.6%) | 33<br>( 7.8%) | 35<br>(8.3%) | 6<br>(1.4%) | 6<br>(1.4%) | 12<br>( 2.8%) | 1<br>(0.002%) |

Notes: i) Items with difficulty or delta indices above 0.75 or Scheuneman indices above the 0.05 critical level for a chi-square distribution with 4 degrees of freedom were classified as biased.

ii) Number of biased items in both samples refers to the number of items classified as biased in both sample one and in sample two.

iii) Overall percentages of biased items are shown in parentheses.

iv) The agreement of classification across samples was evaluated using chi-square tests of independence with Yates' correction. The statistics were 4.70 for difficulty, 3.32 for delta, and 0.66 for Scheuneman. Only the test for the difficulty index surpassed the 0.05 critical value of the chi-square distribution with 1 degree of freedom.

Table 7

Number of Biased Items for Sex

| Test | Number of Items | Index | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Difficulty | | | Delta | | | Scheuneman | | |
| | | Number of Biased Items | | | | | | | | |
| | | Sample One | Sample Two | Both Samples | Sample One | Sample Two | Both Samples | Sample One | Sample Two | Both Samples |
| Vocabulary | 39 | 3 | 3 | 1 | 3 | 2 | 1 | 3 | 0 | 0 |
| Reading | 54 | 3 | 2 | 1 | 3 | 2 | 1 | 0 | 0 | 0 |
| Spelling | 40 | 1 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| Capitalization | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Punctuation | 30 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 0 | 0 |
| Language Usage | 30 | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Visual Materials | 46 | 2 | 3 | 1 | 2 | 3 | 1 | 0 | 2 | 0 |
| Reference Materials | 45 | 2 | 1 | 0 | 2 | 0 | 0 | 2 | 0 | 0 |
| Math Concepts | 37 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| Math Problem Solving | 27 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Math Computation | 45 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 |
| Overall | 423 | 15 (3.5%) | 15 (3.5%) | 3 (0.7%) | 14 (3.3%) | 13 (3.1%) | 3 (0.7%) | 9 (2.1%) | 4 (0.9%) | 0 (0.0%) |

Notes: i) Items with difficulty or delta indices above 0.75 or Scheuneman indices above the 0.05 critical level for a chi-square distribution with 4 degrees of freedom were classified as biased.

ii) Number of biased items in both samples refers to the number of items classified as biased in both sample one and in sample two.

iii) Overall percentages of biased items are shown in parentheses.

iv) The agreement of classification across samples was evaluated using chi-square tests of independence with Yates' correction. The statistics were 7.86 for difficulty, 10.72 for delta, and 4.11 for Scheuneman. Each statistic surpassed the .05 critical value. However, for the Scheuneman statistic this occurred because less than chance agreement was observed.

# Appendix

For the sake of completeness, the reliabilities of signed versions of all but the Scheuneman index were calculated. The reliabilities for signed versions of the difficulty, delta, biserial, and point biserial indices were calculated as described in the paper except that the absolute value of the difference was not taken. The signed 3-parameter index is the overall index described in Linn and Harnisch (1981).

Tables A1 and A2 present the reliabilities. Table A1 corresponds to Table 2 and Table A2 corresponds to Table 3 in the text. Although the signed indices have somewhat greater reliabilities than the unsigned indices, the reliabilities are still consistent with the conclusions stated in the text.

Table A1

Reliability of Signed Item Bias Indices for Race

| Test | Number of Items | Bias Index | | | | |
|---|---|---|---|---|---|---|
| | | Difficulty | Delta | Biserial | Point Biserial | 3-Parameter[a] |
| Vocabulary | 39 | .57* | .46* | .33* | .52* | -.21 |
| Reading | 54 | .52* | .49* | .11 | .23 | |
| Spelling | 40 | .34* | .30 | .17 | .19 | |
| Capitalization | 30 | .28 | .35* | .48* | .60* | |
| Punctuation | 30 | .72* | .70* | .23 | .29 | |
| Language Usage | 30 | .63* | .62* | .44* | .59* | -.34 |
| Visual Materials | 46 | .70* | .63* | .08 | .18 | |
| Reference Materials | 45 | .23 | .15 | .01 | .15 | |
| Math Concepts | 37 | .64* | .60* | .13 | .15 | |
| Math Problem Solving | 27 | .55* | .45* | .44* | .53* | |
| Math Computation | 45 | .38 | .35 | -.02 | .04 | |
| Median | | .55 | .46 | .17 | .23 | -- |

* p < .05

[a] Index was computed only for those tests with values in this column.

Table A2

Reliability of Signed Item Bias Indices for Sex

| Test | Number of Items | Bias Index | | | | |
|---|---|---|---|---|---|---|
| | | Difficulty | Delta | Biserial | Point Biserial | 3-Parameter[a] |
| Vocabulary | 39 | .38* | .39* | .02 | .00 | -.16 |
| Reading | 54 | .46* | .45* | .19 | .20 | |
| Spelling | 40 | .16 | .09 | .25 | .21 | |
| Capitalization | 30 | .11 | .07 | -.05 | .08 | |
| Punctuation | 30 | .13 | .07 | -.17 | -.16 | |
| Language Usage | 30 | -.04 | -.03 | .08 | .08 | -.15 |
| Visual Materials | 46 | .46* | .43* | .01 | .06 | |
| Reference Materials | 45 | .38* | .37* | .00 | .15 | |
| Math Concepts | 37 | .38* | .36* | .22 | .18 | |
| Math Problem Solving | 27 | .02 | .02 | -.11 | -.11 | |
| Math Computation | 45 | .38* | .34* | .21 | .22 | |
| Median | | .38 | .34 | .02 | .08 | -- |

* p < .05

[a] Index was computed only for those tests with values in this column.